ABSTRACT

After prestoring first character strings that occur frequently in words of languages and second character strings that are atypical therein, a device for automatically identifying the language of a text from a plurality of languages extracts words from the text and constructs all of the character strings contained in each extracted word. Each string in an extracted word is compared to the first and second strings of a particular language. If the word contains a first string, a score of the language is increased by a coefficient depending in particular on the position of the first string in the word. If the word contains a second string, the score is decreased by a coefficient associated with the second string. The highest of the scores corresponding to the predetermined languages identifies the language of the text.